LOW-RANK VARIANCE ESTIMATION IN LARGE-SCALE GMRF MODELS

Dmitry M. Malioutov, Jason K. Johnson, and Alan S. Willsky

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge, MA 02139, USA

ABSTRACT

We consider the problem of variance estimation in large-scale Gauss-Markov random field (GMRF) models. While approximate mean estimates can be obtained efficiently for sparse GMRFs of very large size, computing the variances is a challenging problem. We propose a simple rank-reduced method which exploits the graph structure and the correlation length in the model to compute approximate variances with linear complexity in the number of nodes. The method has a separation length parameter trading off complexity versus estimation accuracy. For models with bounded correlation length, we efficiently compute provably accurate variance estimates.

1. INTRODUCTION

We address the problem of estimation in large-scale Gauss-Markov random field (GMRF) models [1]. GMRFs are multivariate jointly Gaussian distributions defined on graphs. The nodes of the graph denote random variables and the edges indicate statistical dependencies between variables. GMRFs can be viewed as generalizations of Markov chains to arbitrary undirected graphs. The Markov property for general graphs (including chains) is that given its neighbors, any node is independent of the rest of the variables in the model. Markovianity has a very transparent manifestation in GMRFs: the inverse covariance $J = P^{-1}$ is sparse according to the graph. An element J_{ij} is non-zero only if the edge $\{i, j\}$ belongs to the edge set of the graph. GMRFs arise in a wide variety of important practical applications from fields including computer vision, spatial statistics (e.g. geostatistics and oceanography), and spline interpolation among others [1]. A prototypical application is surface interpolation based on a set of sparse irregular noisy measurements [3].

Estimating means and variances of hidden variables based on observations in GMRFs can be done by matrix inversion. However, for large-scale problems with millions of variables, exact algorithms such as Gaussian elimination (with $O(N^3)$ complexity in the number of variables, N) are intractable. Approximate mean estimates can be computed with $O(N)$ complexity for sparse graphs by iterative solvers such as conjugate gradients or multigrid techniques. However conditional means are most useful if the confidence in the estimates (i.e. variance) is also known.

Variances can be computed by more efficient versions of Gaussian elimination (based on junction trees) which take the graph structure into account and reduce the complexity to being cubic in the "tree-width" of the graph. For square lattice models, the tree-width is equal to the width of the graph, so the complexity reduces to $O(N^{3/2})$. Despite being a great improvement, this is still intractable for large models; in addition, the implementation of such algorithms is rather complex. Approximate methods such as belief propagation (BP) [2] have linear complexity in N per iteration, but they are not guaranteed to converge, and convergence may be very slow for large problems. Even in case of convergence, the BP variance estimates may be rather poor. For stationary models, efficient FIR approximations can be used for both means and variances, but for non-stationary problems such as interpolation from sparse noisy measurements, these approximations do not apply.

In this paper we propose a rank-reduced method to compute approximate variances with linear complexity in the number of nodes for large-scale sparse GMRFs. In addition to its simplicity and speed the method gives unbiased estimates, and for models with bounded correlation length we prove that the variance estimates are very accurate. In Section 2 we discuss estimation with GMRF models. We describe and analyze our rank-reduced method to compute variance estimates in lattice models and arbitrary graphs in Section 3. We apply our method to sea surface height altimetry data in Section 4.

2. ESTIMATION IN GMRF MODELS

A GMRF model is defined by a graph $\mathcal{G} = (V, \mathcal{E})$ with vertices V and edges $\mathcal{E} \subset \binom{V}{2}$, i.e., some set of two-element subsets of V, and a collection of jointly Gaussian random variables $x = (x_i, i \in V)$ with probability density given in *information form*:

$$p(x) \propto \exp\{-\frac{1}{2}x'Jx + h'x\} \tag{1}$$

This research was supported by the Air Force Office of Scientific Research under Grant FA9550-04-1, the Army Research Office under Grant W911NF-05-1-0207, and by a grant from MIT Lincoln Laboratory.

The matrix J is called the information matrix, and it is symmetric positive definite ($J \succ 0$) and sparse so as to respect the graph $\mathcal{G}$: if $\{i,j\} \notin \mathcal{E}$ then $J_{ij} = 0$. We call h the potential vector. These quantities are directly related to the usual parameterization of Gaussian densities in terms of the mean $\mu \equiv \mathbb{E}\{x\}$ and the covariance matrix $P \equiv \mathbb{E}\{(x-\mu)(x-\mu)'\}$:

$$\mu = J^{-1}h \quad \text{and} \quad P = J^{-1} \tag{2}$$

Let $A \subset V$. Define x_A to be the vector $(x_i | i \in A)$ corresponding to the variables in A. Let $V \backslash A$ denote the complement of A. We use a shorthand $V \backslash i \triangleq V \backslash \{i\}$. Let $N(i) = \{j | \{i,j\} \in \mathcal{E}\}$ denote the *neighbors* of i in the graph. Then the GMRF model satisfies the Markov property:

$$p(x_i | x_{V\backslash i}) = p(x_i | x_{N(i)}), \quad \forall i \tag{3}$$

We add observations y_i of the hidden variables x_i, with Gaussian $p(y_i|x_i)$. Assume that y_i is independent of x_j and other y_j for $j \neq i$: $p(y|x) = \prod_{i \in V} p(y_i|x_i)$. The posterior, $p(x|y) \propto p(y|x)p(x)$ is a function of x (y is observed and does not change) that has the same form as $p(x)$. Introducing the observations modifies the information parameters h and the diagonal of J. An estimate of x is specified by the marginals of $p(x|y)$, the conditional means and variances of x given y.

Example GMRF: thin plate model. Consider the thin plate model commonly used for data interpolation:

$$p(x|y) \propto \exp\left(-\alpha \sum_{i \in V} (x_i - \frac{1}{N_i} \sum_{j \in N(i)} x_j)^2 - \beta L(x_i, y_i)\right) \tag{4}$$

The first term is the prior enforcing smoothness, i.e. x_i should be close to the mean of its neighbors $\{x_j, j \in N(i)\}$, with $N_i = |N(i)|$. The second term is the data term, $L(x_i, y_i) = \sum_{i \in V} (x_i - y_i)^2$, which makes the estimates consistent with the observations. From this specification, the information form parameters J and h are readily obtained[1].

Given a model in information form, it is of interest to estimate the (conditional) means μ and the variances P_{ii} for all x_i. Using matrix inversion for this task is intractable for large-scale models. Approximate mean estimates can be efficiently obtained by a sparse linear system solver, since J is sparse. However, as we described in Section 1, no simple tractable methods exist to compute the variances in general graphs[2]. An estimate of the means is much less insightful when the confidence in it (the variance) is unknown. Next we describe a family of rank-reduced methods to provide efficient variance estimates in GMRFs.

[1]The thin plate model is Markov on a modification of the graph defining the thin plate model: edges between nodes two steps away have to be added.

[2]A new method that has tractable computation of approximate variances is recursive cavity modeling (RCM) [3]. However, it employs the machinery of information geometry and is quite involved to implement. In contrast, the method of this paper is simple and also provides theoretical guarantees of quality for the variance estimates.

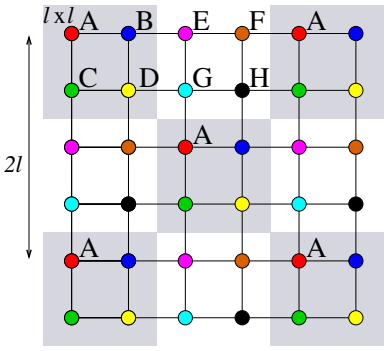

Fig. 1. Local 2x2 regions for square lattice. Odd blocks are shaded, and even ones are transparent. Colors: $\{A, .., H\}$.

3. REDUCED-RANK VARIANCE CALCULATION

For sparse graphs where the number of edges is far less than in the full graph, $|\mathcal{E}| \ll |V|^2$, efficient mean estimates to a given tolerance can be obtained by a sparse iterative solver: $J\mu = h$. If the number of neighbors for every node is bounded by a small constant (e.g. 4 for lattices), then methods with $O(N)$ (where $N = |V|$) complexity may be used for the means.

Sparse iterative solvers can compute the variances as well: let $e_i \in \mathbb{R}^N$ be the i-th standard basis vector, then the i-th column of P can be obtained as $JP_i = e_i$. This has to be done N times, at each node: $JP = [e_1, ..., e_N] = I$, leading to an intractable complexity of $O(N^2)$. Note, that in each vector P_i we are interested in the i-th element only, P_{ii}.

The computation $\text{diag}(P) = \text{diag}(J^{-1}I)$ is costly. We would like to design a low-rank matrix BB', with $B \in \mathbb{R}^{N \times M}$ and $M \ll N$, and use it instead of I. Let all rows b_i of B have unit norm: $b_i' b_i = 1$. Consider the quantity $\text{diag}(J^{-1}(BB'))$ which is tractable to compute. Then:

$$\hat{P}_{ii} \triangleq [J^{-1}(BB')]_{ii} = P_{ii} + \sum_{i \neq j} P_{ij}\, b_i' b_j \tag{5}$$

To force $\hat{P}_{ii}$ to be accurate estimates of the variances we need $P_{ij}\, b_i' b_j$ to be nearly zero for all pairs of nodes. For a wide class of models, correlation P_{ij} decays with distance from i to j. We define the correlation length d as the distance after which P_{ij} decreases to a small fraction[3] of P_{ii}. For nodes which are far away in the graph (further than the correlation length), P_{ij} is negligible. For nodes which are nearby, the terms b_i and b_j have to be (nearly) orthogonal. Thus we are led to the problem of designing an overcomplete basis $\{b_i \in \mathbb{R}^M\}$ which is nearly orthogonal with respect to a graph $\mathcal{G}$. We describe such a construction for rectangular lattices first, and in Section 3.3 we extend it to arbitrary graphs.

Constructing B in rectangular lattices. Consider a rectangular $K \times L$ lattice (with $N = KL$ total variables). Partition the lattice into small blocks of size $l \times l$ in a checker-board

[3]If d depends on i, then we take the maximum over i.

pattern, see Figure 1. We partition the nodes into $M = 2l^2$ classes (we call them colors), with l^2 in even blocks and l^2 in odd blocks. The minimum distance (shortest path in the graph) between nodes of the same color is $2l$. We choose the vectors b_i and b_j to be orthogonal if nodes i and j have different colors: for each class $c \in \{1, .., M\}$ we allocate a 1-dimensional subspace, e.g. span(e_c), where $e_c \in \mathbb{R}^M$ (c-th standard basis vectors for $\mathbb{R}^M$). Thus $P_{ij}\ b_i' b_j = 0$ for nodes of different color. We randomly assign directions along e_c to each node in the class: $b_i = s_i e_c$, where $s_i \in \{-1, 1\}$ are independent. By making the separation length l comparable to the correlation length in the model, we satisfy the condition $P_{ij}\ b_i' b_j \approx 0$ for nodes i and j of the same color.

Algorithm summary:
1) Construct $B \in \mathbb{R}^{N \times M}$.
2) For each color $c = 1, .., M$:
   i) Let b equal the c-th column of B. Solve $Jr = b$.
   ii) For each node i with color c, $\hat{P}_{ii} = s_i r_i$.

3.1. Properties of the estimate $\hat{P}$

Next, we show that our reduced rank computation is tractable, unbiased, and the error in the estimates can be driven to zero by taking the separation length large enough. In terms of computational complexity, the approximate solution of $JR = B$ has complexity $O(MN)$ using iterative solvers. The step of post-multiplication by B', i.e. $\hat{P}_{ii} = [RB']_{ii}$ requires MN operations (we only need the diagonal). Hence, for fixed number of colors M, the complexity is linear in N.

Unbiased. The estimates $\hat{P}_{ii}$ are unbiased. $E[\hat{P}_{ii}] = P_{ii} + \sum_{j \neq i} P_{ij} E[b_i' b_j]$. But, $E[b_i' b_j] = 0$ both for nodes of different color (due to orthogonality), and for nodes of the same color c, since $b_i = s_i e_c$, with independent s_i: $E[s_i s_j] = 0$.

Variance of the estimates. Suppose that the correlations P_{ij} fall off exponentially with the distance $d(i, j)$ between i and j, i.e. $P_{ij} \leq A\, \alpha^{d(i,j)}$, with $0 \leq \alpha < 1$. This is true for a wide class of models including Markov models on bipartite graphs. Now, $\text{Var}(\hat{P}_{ii}) = E[(\hat{P}_{ii} - P_{ii})^2] = E[(\sum_{j \neq i} P_{ij} b_i' b_j)^2]$. Let $\mathcal{C}(i)$ be the set of nodes of the same color. For $j \notin \mathcal{C}(i)$, $b_i' b_j = 0$. Using iterated expectations, we have:

$$\begin{aligned}
\text{Var}(\hat{P}_{ii}) &= E\left\{ E\left\{ \left(\sum_{j \in \mathcal{C}(i) \setminus i} P_{ij} b_i' b_j \right)^2 \mid b_i \right\} \right\} \\
&= E\left\{ \sum_{j \in \mathcal{C}(i) \setminus i} P_{ij}^2\ E\left\{ (b_i' b_j)^2 \mid b_i \right\} \right\} \qquad (6) \\
&= \sum_{j \in \mathcal{C}(i) \setminus i} P_{ij}^2\ E\left\{ (b_i' b_j)^2 \right\} = \sum_{j \in \mathcal{C}(i) \setminus i} P_{ij}^2
\end{aligned}$$

The equation in the second line holds because variance of a sum of conditionally independent random variables equals the sum of the variances. The equation in the third line uses linearity of expectation, and that $(b_i' b_j)^2 = 1, \forall j \in \mathcal{C}(i)$.

In a lattice model with our construction, the number of nodes of a given color which are $(2l)n$ steps away is $8n$ (all the distances between nodes of the same color are integer multiples of $2l$). Using the exponential decay bound, for nodes j with $d(i, j) = 2nl$, $P_{ij} = A\, \alpha^{2nl}$. Hence,

$$\sum_{j \in \mathcal{C}(i) \setminus i} P_{ij}^2 \leq \sum_{n=1}^{\infty} 8nA^2\, \alpha^{4nl} = 8A^2 \frac{\alpha^{4l}}{(1 - \alpha^{4l})^2} \qquad (7)$$

We have used the following series: $\sum_{n=1}^{\infty} n\beta^n = \frac{\beta}{(1-\beta)^2}$. Thus, $\text{Var}(\hat{P}_{ii}) \leq 8A^2 \frac{\alpha^{4l}}{(1-\alpha^{4l})^2}$. Since, $|\alpha| < 1$, we can choose l large enough such that the variance of the estimate is below any desired threshold. In practice, l should be chosen to be comparable to the correlation length of the model.

The above error bounds and the fact that our method is unbiased make it especially attractive compared to a windowing method[4], where such guarantees are not available.

3.2. Efficient preconditioners

Our reduced-rank method estimates the variances by solving M systems of linear equations, all sharing the same linear operator, J. It is thus beneficial to design a good preconditioner for J to improve the convergence speed of the sparse iterative solvers such as Richardson iterations or conjugate gradients [4]. A preconditioner for a linear system $Jx = y$ is a matrix Q such that the system $QJx = Qy$ is easier to solve (the matrix QJ is better conditioned than J). Ideally, $Q = J^{-1}$, making the transformed linear system trivial. However, applying Q in this case is as hard as solving the original problem. The inverse of the diagonal or tridiagonal part of J are examples of easily computable preconditioners.

For the lattice GMRF model a very efficient set of preconditioners based on embedded trees (ET) has been developed in [4]. The idea is that for models with a tree-structured graph $\mathcal{G}$, inversion of the information matrix J is extremely efficient - it can be done in $O(N)$ operations. Hence for general graphs $\mathcal{G}$, [4] uses spanning trees $T \subset \mathcal{G}$ with $Q = J_T^{-1}$. We use a variant[5] of ET in experiments in Section 4.

3.3. Node class definition for arbitrary graphs

We propose to extend our coloring construction from lattices to arbitrary graphs. For an arbitrary graph $\mathcal{G}$ we would like to divide the vertices into M disjoint subsets, such that any two nodes belonging to the same subset are separated by at least l

[4] At each node an exact estimate over a small window is produced, ignoring the rest of the graph.

[5] We use alternating sets of narrow induced subgraphs. In Richardson iterations this guarantees convergence by equivalence with block Gauss-Seidel.

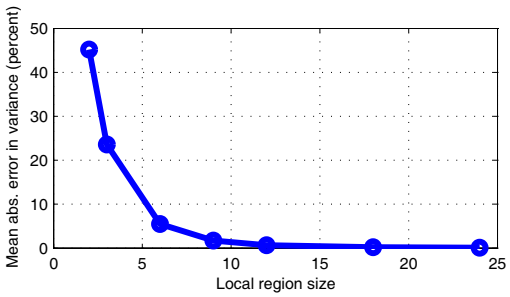

Fig. 2. Errors in variance estimates vs. separation length l.

steps. We approach this problem by approximate graph coloring in an augmented graph. For a desired minimum distance l we connect every pair of nodes that are within l steps away by an edge. Assigning M colors such that no neighbors in the augmented graph share the same color would solve our problem. This is a graph coloring problem, and it is known to be hard. We settle for an approximate solution that allows a few violations and uses more than the minimum required number of colors. Fast techniques for approximate graph coloring include belief propagation (in max-product form), and eigen-decomposition-based methods [5]. After defining the node classes, variances can be estimated in the same way as for lattices. The method is expected to be useful for graphs which have the topology where for any given node only a few nodes are "near", and most nodes are "far".

4. RESULTS

We now apply the approximate variance calculation method to the ocean surface height data collected along the tracks of Jason-1 satellite[6]. We compute the variance of the estimates over the Pacific ocean region. The data is sparse and highly irregular. We use the thin-plate model for the data. First, we select a moderate resolution for which exact estimation of variances using the junction tree method is feasible, and compare the approximate answers from our reduced-rank method to the exact variances. The size of the model is 288×432, with 0.325 degree spacing in both latitude and longitude. The plot of mean absolute error in variances versus the separation length l appears in Figure 2. It can be seen that the errors rapidly decrease to zero, and for region sizes greater than $l = 12$, very accurate variances are obtained. We note that Gaussian belief propagation diverged for this application.

Next we increase the resolution to 0.129 degrees, resulting in a grid of size 720×1080. Estimating the variance in a model of this size is beyond what is practical with exact methods on a single workstation. We use our approximate variance calculation method. The separation length l is increased to 30×30 to account for the increase in correlation length when resolution is increased. The resulting estimate appears

[6]This altimetry data is available from the Jet Propulsion Laboratory http://www.jpl.nasa.gov. It is over a ten day period beginning 12/1/2004.

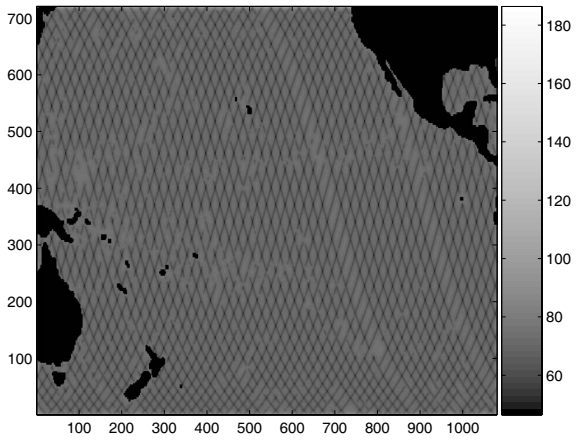

Fig. 3. Uncertainty estimates of Pacific ocean surface height based on measurements along satellite tracks, 720x1080 grid.

in Figure 3. The regions over land are ignored (in black). The variances are lowest near the measurements (along the tracks) as expected.

5. CONCLUSION

We presented a scalable tractable approach to calculate approximate variances in large scale GMRF estimation problems. We justified the approach both theoretically and with experiments on satellite altimetry data. An important direction for further work is to develop a multiscale version of the approach to decrease the correlation length at the finest scale, and make it viable for a wider class of models.

6. REFERENCES

[1] H. Rue and L. Held, *Gaussian Markov Random Fields Theory and Applications*, Chapman and Hall, CRC, 2005.

[2] J. K Johnson, D. M. Malioutov, and A. S. Willsky, "Walk-sum interpretation and analysis of Gaussian belief propagation," in *NIPS*, 2005.

[3] J.K. Johnson and A.S. Willsky, "A recursive model-reduction method for approximate inference in Gaussian Markov random fields," *IEEE Trans. Imag. Proc.*, 2005 (in review).

[4] E. Sudderth, M. J. Wainwright, and A. S. Willsky, "Embedded trees: Estimation of Gaussian processes on graphs with cycles," *IEEE Trans. Signal Proc.*, vol. 52, pp. 3136–3150, Nov. 2004.

[5] B. Aspvall and J. R. Gilbert, "Graph coloring using eigenvalue decomposition," *SIAM J. Alg. Disc. Meth.*, vol. 5, pp. 526–538, 1984.